CLONAGEM DE VOZ COM INTELIGÊNCIA ARTIFICIAL PARA ROBÔS EDUCATIVOS: DESAFIOS NA PRODUÇÃO DE ÁUDIOS PARA ALFABETIZAÇÃO E EXPRESSÃO EMOCIONAL

Maria Vittória Sousa Rodrigues Palma
Universidade Federal de Mato Grosso (UFMT)
vitlia40@gmail.com - http://lattes.cnpq.br/9190185223212883

Thais Reggina Kempner
Universidade Federal de Mato Grosso (UFMT)
thaisrgk@gmail.com - http://lattes.cnpq.br/5630808549813002

Luciana C. L. de Faria Borges
Universidade Federal de Mato Grosso (UFMT)
lucianafariaborges@gmail.com - http://lattes.cnpq.br/8434103815800687

Eunice P. dos Santos Nunes
Universidade Federal de Mato Grosso (UFMT)
eunice@ufmt.br - http://lattes.cnpq.br/0992498795960159

Resumo: Este trabalho explora o potencial da clonagem de voz por Inteligência Artificial (IA) aplicada ao robô BITI, uma ferramenta voltada à alfabetização de crianças neurodivergentes pelo método fônico. Diante da limitação na expansão dos áudios originais de uma criança, foram investigadas alternativas capazes de gerar vozes com naturalidade, precisão fonética e expressão emocional, utilizando a plataforma Eleven Labs em conjunto com o software Audacity. Os testes envolveram frases com carga emocional e ajustes de parâmetros diversos, revelando desafios na reprodução de emoções e de fonemas isolados. Os resultados apontam que a combinação entre IA e supervisão humana pode viabilizar a criação de áudios adequados ao contexto educacional, abrindo caminhos promissores para práticas pedagógicas mais acessíveis e personalizadas.
Palavras-chave: Clonagem de voz por IA. Tecnologia assistiva. Expressividade emocional. Autenticidade vocal.

Abstract: *This study explores the potential of voice cloning through artificial intelligence applied to the BITI robot, a tool designed to support the literacy of neurodivergent children using the phonics method. Faced with limitations in expanding a child's original audio recordings, alternative solutions were investigated to generate speech with naturalness, phonetic accuracy, and emotional expression, using the Eleven Labs platform in combination with Audacity software. Tests involving emotionally charged phrases and parameter adjustments revealed challenges in reproducing emotions and isolated phonemes. The findings suggest that combining AI with human supervision can enable the creation of suitable audio content for educational purposes, opening promising paths toward more accessible and personalized pedagogical practices.*
Keywords: *AI voice cloning. Assistive technology. Emotional expressiveness. Vocal authenticity.*

DOI: 10.29327/2686685.1.4-26

INTRODUÇÃO

A alfabetização é um marco essencial no desenvolvimento infantil, sendo fundamental para a interação com o mundo por meio da linguagem escrita. Para crianças atípicas, cujos padrões de desenvolvimento diferem do neurotípico (Abreu, 2006), esse processo é ainda mais complexo. Neste estudo, o foco está especialmente em crianças com Transtorno do Espectro Autista (TEA), Transtorno do Déficit de Atenção com Hiperatividade (TDAH) e Dislexia, que necessitam de abordagens educacionais especializadas, mas que também encontram respaldo em legislações específicas no Brasil.

O conjunto de leis que sustenta essa proposta educacional abrange a Lei Berenice Piana (12.764/2012), a qual assegura direitos fundamentais às pessoas com autismo, e a Lei 14.254/2021, a qual estabelece diretrizes para o atendimento educacional de estudantes com TDAH e Dislexia (Brasil, 2012, 2021). Esses instrumentos legais não somente legitimam, mas demandam a criação de estratégias pedagógicas inovadoras, capazes de superar as barreiras impostas por dificuldades de comunicação, interação social e processamento sensorial características desses grupos (Bar-shalita; Vatine; Parush, 2008).

Nesse cenário, abordagens tecnológicas têm sido exploradas para apoiar o processo de ensino-aprendizagem, como o uso de robôs sociais como ferramenta de mediação. Estudos têm mostrado que essas tecnologias podem contribuir para aumentar o engajamento e desenvolvimento de habilidades comunicativas em crianças com TEA (Scassellati; Admoni; Mataric, 2012), Dislexia (Rosenberg-Kima *et al.*, 2019) e TDAH (Berrezueta-Guzman *et al.*, 2022), ao proporcionarem ambientes de aprendizagem estruturados, previsíveis e emocionalmente seguros.

Um exemplo é o robô Otto, desenvolvido no projeto FATA (Fábrica de Alta Tecnologia Assistiva), que utiliza a tecnologia RFID (*Radio-Frequency Identification*) em conjunto com recursos de reprodução de áudio para facilitar o ensino de palavras que representam emoções, objetos, frutas (Rebouças *et al.*, 2023; Dias *et al.*, 2023). Ao aproximar um cartão RFID do robô, são reproduzidos áudios que traduzem a imagem contida no cartão, promovendo uma interação dinâmica entre o robô e a criança,

tornando as experiências de aprendizagem mais interativas, expressivas e personalizadas.

Com base nessas experiências, a evolução natural do trabalho levou à concepção do robô BITI (Brinquedo Inclusivo para a Tutoria Infantil), um projeto inovador com o propósito de apoiar no processo de alfabetização por meio do método fônico – abordagem reconhecidamente eficaz para crianças neurodivergentes (Oliveira; Albuquerque, 2021). No entanto, o desenvolvimento do BITI trouxe à tona um desafio técnico relevante: a impossibilidade de reutilizar os áudios originais gravados por um participante de 11 anos, em razão das mudanças vocais naturais decorrentes do crescimento (Rebouças *et al.*, 2023).

Diante disso, a clonagem vocal por Inteligência Artificial (IA) passou a ser investigada como uma alternativa promissora para contornar essas limitações. Apesar dos avanços recentes nessa tecnologia (Pepino *et al.*, 2022; Neekhara *et al.*, 2021), sua aplicação em contextos educacionais sensíveis impõe desafios consideráveis, especialmente no que diz respeito à reprodução precisa de variações prosódicas e expressões emocionais sutis – aspectos fundamentais para a eficácia pedagógica (Chen; Jiang, 2023).

No âmbito do projeto, destaca-se a utilização da plataforma Eleven Labs para geração de áudios sintéticos, dando continuidade a uma iniciativa anteriormente aplicada ao robô Otto (Palma *et al.*, 2024). A preservação do sotaque original mostrou-se essencial, considerando que algumas crianças autistas apresentam hipersensibilidade a variações sonoras durante a reprodução da fala (Gomes *et al.*, 2008). Dessa forma, esta pesquisa concentra-se na análise da capacidade da IA em gerar entonações com expressividade emocional, investigando os principais desafios técnicos envolvidos na criação de uma voz autêntica e na fidelidade dessa reprodução em contextos educacionais com crianças neurodivergentes.

A qualidade dos áudios utilizados nas interações com o robô BITI está vinculada à capacidade de a tecnologia sintetizar vozes com naturalidade e carga emocional convincente – o que se revela essencial para promover uma comunicação fluida, engajadora e emocionalmente notável (Pinheiro *et al.*, 2021), especialmente quando se busca estabelecer vínculos afetivos no processo de ensino-aprendizagem com crianças neurodivergentes.

A relevância deste trabalho é ampliada por sua consonância com o objetivo de Desenvolvimento Sustentável nº 10 da Organização das Nações Unidas, que trata da redução das desigualdades. A proposta dialoga diretamente com políticas públicas de educação inclusiva e com o potencial das tecnologias educacionais para democratizar o acesso a recursos personalizados. Destaca-se, contudo, a importância de assegurar o rigor científico na aplicação dessas tecnologias, a fim de que atendam de forma eficaz às necessidades específicas de crianças neurodivergentes, sem comprometer a qualidade pedagógica.

1. DESENVOLVIMENTO

A aplicação de sistemas de IA na educação tem revolucionado as práticas pedagógicas, particularmente no atendimento a crianças neurodivergentes (Albino *et al.*, 2024). No contexto específico do robô BITI, a IA assume papel central na personalização do ensino, permitindo a criação de áudios personalizados e adaptados às necessidades individuais de aprendizado.

Conforme apontado por Pepino *et al.* (2022) e Neekhara *et al.* (2021), os avanços recentes em síntese vocal e clonagem de voz têm ampliado o alcance das aplicações de IA ultrapassando os limites do entretenimento para oferecer soluções inovadoras em ambientes educacionais. Tais avanços tornam possível a criação de conteúdos auditivos que não apenas transmitem informação, mas o fazem de forma humanizada e responsiva às necessidades emocionais e sensoriais do público infantil.

O uso de IA no projeto BITI foca na geração de áudios que atendam a três critérios essenciais: adequação sensorial, precisão fonética e expressividade emocional. Os áudios devem respeitar a sensibilidade auditiva de crianças com TEA, que frequentemente apresentam hipersensibilidade auditiva (Gomes *et al.*, 2008), garantir a reprodução clara dos fonemas e favorecer a associação entre grafemas e fonemas, conforme o método fônico.

Conforme apontam Shaywitz *et al.* (2003), a dislexia está diretamente relacionada a déficits nesse tipo de processamento, o que reforça a importância de estímulos auditivos precisos para crianças com dificulda-

des de leitura. Já a expressividade emocional é uma característica humana complexa e multifacetada, sendo resultado de modificações fisiológicas que induzem variações nos parâmetros acústicos da fala, como frequência fundamental, intensidade, timbre e ritmo, tornando a reprodução artificial dessas nuances uma das maiores barreiras na clonagem vocal (Crivellaro e Silva, 2012).

Por exemplo, frases como "estou cansado..." foram redigidas com reticências para induzir uma entonação arrastada, evocando cansaço ou apatia. Já expressões como "está MUITO barulhoooo" empregaram letras maiúsculas e prolongamento vocálico como estratégias explícitas para transmitir sensação de incômodo e sobrecarga sensorial. Esses recursos textuais funcionam como guias interpretativos para o modelo de IA orientando a produção da fala de forma mais natural e alinhada ao conteúdo emocional pretendido.

A construção dos áudios foi um processo interativo, com avaliações contínuas da equipe multidisciplinar. Apenas os áudios que atendiam aos critérios de naturalidade, clareza e fidelidade emocional eram validados, sendo os *prompts* reformulados sempre que necessário. Essa abordagem destaca a importância do domínio do *prompting* para alcançar interações mais humanas, especialmente no atendimento às necessidades de crianças com TEA, TDAH e Dislexia. A plataforma Eleven Labs contribuiu nesse processo ao oferecer algoritmos de *deep learning* capazes de gerar vozes com alta fidelidade expressiva.

Contudo, o processo de clonagem vocal para fins educacionais envolve etapas complexas que vão além da mera reprodução de palavras. Como destacam Chen e Jiang (2023), a síntese de voz para aplicações pedagógicas exige atenção especial à articulação de fonemas isolados - requisito essencial para o método fônico. O desafio técnico reside na capacidade dos sistemas atuais de IA em reproduzir com precisão os sons fonéticos básicos, especialmente quando aplicados a línguas com características fonológicas específicas como o português brasileiro.

A análise comparativa entre áudios originais e clonados mostra que, apesar dos avanços tecnológicos na reprodução de frases completas, ainda existem desafios na geração de fonemas isolados com articulação clara, transições suaves entre consoantes e vogais, e padrões entoacionais que expressem adequadamente diferentes estados emocionais (Neekhara

et al., 2021). Essas limitações são especialmente importantes na alfabetização fônica, na qual a precisão na reprodução dos sons das letras é fundamental para o sucesso do método (Marttins, 2022).

A escolha do método fônico para a alfabetização de crianças neurodivergentes está fundamentada na literatura, como mostram Oliveira e Albuquerque (2021), sendo particularmente eficaz para crianças com TEA por trabalhar diretamente a relação grafema-fonema, reduzindo a necessidade de habilidades de abstração que frequentemente são difíceis para essa população.

O método fônico impõe aos sistemas de síntese vocal a exigência de precisão fonética rigorosa, com cada som sendo reproduzido de forma fiel, a fim de evitar ambiguidades no processo de aprendizagem. Além disso, requer consistência entoacional, preservando os padrões fonéticos essenciais mesmo diante da necessária variação prosódica associada à expressão emocional (Sebra; Dias, 2011).

A pesquisa de Fragoso *et al.* (2013) expande essa discussão ao indicar que crianças com TDAH se beneficiam da combinação do método fônico com pistas auditivas distintas. Nesse sentido, a capacidade do robô BITI de modular a expressão emocional sem prejudicar a clareza fonética representa um progresso potencialmente significativo para a alfabetização inclusiva.

A implementação bem-sucedida da clonagem vocal no BITI depende da superação de desafios que se situam na intersecção entre tecnologia e pedagogia. Por um lado, os requisitos técnicos envolvem o refinamento de algoritmos para capturar nuances prosódicas essenciais à comunicação educacional efetiva. Por outro lado, a dimensão pedagógica exige que esses recursos tecnológicos estejam alinhados com os princípios do desenvolvimento infantil e das necessidades específicas de aprendizes neurodivergentes.

Os estudos de caso analisados por Feil-Seifer e Mataric (2009) com robôs sociais sugerem que a combinação entre consistência comportamental e variabilidade emocional moderada pode otimizar os resultados educacionais. Essa aparente contradição – que exige simultaneamente previsibilidade e expressividade – constitui o núcleo do desafio enfrentado pelo projeto BITI em sua proposta de integrar clonagem vocal e método fônico.

2. METODOLOGIA

A presente pesquisa configura-se como um estudo de caso sobre a clonagem de voz infantil por meio de IA, com ênfase na produção de materiais sonoros voltados ao robô educacional BITI, destinado a crianças neuroatípicas. O projeto foi aprovado previamente pelo Comitê de Ética em Pesquisa com Seres Humanos da Universidade Federal de Mato Grosso (UFMT), em conformidade com os preceitos éticos exigidos para estudos envolvendo menores de idade.

A primeira etapa consistiu na coleta de 170 gravações de áudio realizadas em estúdio, utilizando a voz de uma criança de 11 anos. Esses áudios foram empregados na construção de um acervo inicial de conteúdos interativos para o robô. No entanto, a mudança natural na voz da criança – decorrente do desenvolvimento fisiológico – impossibilitou a continuidade das gravações com a mesma qualidade e timbre vocal. Diante dessa limitação, a equipe de pesquisa optou por empregar técnicas de clonagem de voz baseadas em IA como estratégia alternativa para garantir a consistência dos áudios e ampliar o repertório sonoro do robô sem a necessidade de novas gravações presenciais.

Para embasar teoricamente o estudo e orientar as decisões técnicas, foi realizado um levantamento bibliográfico nas bases IEEE Xplore, Google Scholar e Scopus. As estratégias de busca incluíram os seguintes descritores: *"Artificial Intelligence" AND "voice cloning" AND "autism"*, *"voice synthesis" AND "robot interaction"*, e *"human-robot interaction" AND "autistic children"*. Essa revisão teve como foco identificar soluções baseadas em IA capazes de produzir vozes clonadas com qualidade suficiente para otimizar a interação entre robôs e crianças com TEA, mantendo a autenticidade vocal e a coerência emocional.

Com base nos dados obtidos, foram analisadas três plataformas especializadas em clonagem de voz: Eleven Labs (https://elevenlabs.io/), Play HT (https://play.ht) e VoiceIA (https://voice.ai/). A escolha recaiu sobre a Eleven Labs devido a seus algoritmos avançados, capazes de reproduzir fielmente padrões de fala, entonação e expressividade. Outros fatores decisivos incluíram a boa relação custo-benefício e os relatos positivos de aplicação em contextos educacionais e terapêuticos.

Para o treinamento da IA, foram utilizadas as 170 frases gravadas previamente pela criança de 11 anos em ambiente controlado. Os áudios foram processados e aprimorados, com especial atenção ao volume e à frequência, considerando que algumas crianças autistas apresentam hipersensibilidade durante a reprodução sonora (Gomes *et al.*, 2008). Esse cuidado permitiu uma clonagem vocal precisa e compatível com as características expressivas da voz original (Casanova *et al.*, 2023).

Diante da necessidade de simplificar alguns áudios do estudo, optou-se pelo *software* de edição de áudio gratuito e de código aberto Audacity, escolhido por seus recursos de edição e ampla compatibilidade com diversas plataformas (Jaworski; Thibeault, 2011). Essa ferramenta permite editar, processar e aplicar efeitos aos áudios existentes, além de auxiliar na manipulação dos novos áudios gerados pela Eleven Labs. Entre suas funcionalidades, destaca-se a possibilidade de modificar a taxa de amostragem (em Hz), ajustar a velocidade de reprodução sem comprometer o tom e alterar a frequência sonora. Tais recursos mostraram-se essenciais para garantir uma experiência auditiva confortável, sobretudo em função da hipersensibilidade auditiva frequentemente observada em crianças autistas.

Para superar o desafio de incorporar emoção e personalidade aos áudios – uma limitação observada na Eleven Labs – foram implementados testes contínuos para comparar e refinar os resultados obtidos. Na geração de áudios com entonação emocional realista, aplicaram-se técnicas específicas de construção de *prompts*, orientando a IA na modulação vocal conforme a emoção pretendida. A formulação desses comandos textuais envolveu o uso intencional de recursos linguísticos como pontuação expressiva (reticências, exclamações), prolongamentos vocálicos, uso de letras maiúsculas para indicar ênfase e escolhas lexicais específicas. Tais elementos funcionam como indicadores prosódicos, capazes de influenciar a interpretação da IA e, consequentemente, a expressividade do áudio gerado. Essa metodologia visa garantir que a voz do robô BITI promova interação e engajamento verbal de maneira eficaz e natural.

3. RESULTADOS E DISCUSSÃO

Durante os testes de clonagem vocal com a plataforma Eleven Labs, observou-se uma discrepância entre os áudios originais e os gerados artificialmente. Apesar de múltiplas tentativas e regenerações, apenas uma fração dos áudios sintetizados foi aprovada pela equipe técnica, evidenciando a necessidade de um processo de filtragem criterioso. Embora a ferramenta venha apresentando melhorias graduais na naturalidade da voz, os avanços em direção à adequação ainda ocorrem de forma lenta. Ressalta-se que a plataforma tem sido utilizada e alimentada há aproximadamente um ano e meio – um período relativamente curto diante da complexidade envolvida na replicação fidedigna dos traços fonéticos e emocionais da fala infantil.

A avaliação dos áudios gerados foi baseada principalmente em dois critérios: naturalidade, entendida como o quanto a voz sintetizada se aproxima da fala humana em fluidez e entonação; e inteligibilidade, relacionada à clareza na articulação dos sons e palavras. Observou-se que a plataforma utilizada, Eleven Labs, opera a partir de parâmetros técnicos predefinidos – velocidade, estabilidade, similaridade e intensidade emocional – que funcionam de forma relativamente autônoma e não captam o tom emocional sugerido pelo *prompt* textual. Isso ocorre porque a IA não interpreta semanticamente o conteúdo escrito; sua resposta vocal é guiada pelas configurações técnicas e pelo histórico de *feedback*.

Essa limitação torna o processo de geração emocionalmente adequado imprevisível: as variações prosódicas e afetivas mudam a cada regeneração, e a plataforma busca ajustar-se com base na aceitação ou rejeição das versões anteriores. Apesar dessa limitação, para vocábulos cotidianos, como nomes de alimentos – por exemplo, "alface" – a reprodução gerada pela IA[1] é bastante satisfatória, com qualidade próxima à do áudio original[2].

No entanto, em frases que exigem maior carga emocional, a presença de entonação expressiva apresentou-se mais desafiadora. Com o objetivo de testar a expressividade emocional, foram selecionadas três expressões: "Estou feliz", "Estou triste" e "Estou com sono". No experi-

[1] https://drive.google.com/file/d/1O1IhoABAOu-LLJ6vqKzJUW2p2ZwlMe4E/view
[2] https://drive.google.com/file/d/1WX6AZljBPMiW4XPxiJe0XuN9DO9x8PRP/view

mento voltado à reprodução da emoção associada à frase "Estou feliz", originalmente gravada em estúdio com entonação expressiva[3], buscou-se avaliar a capacidade da plataforma de IA em replicar o tom de felicidade por meio de variações simples no *prompt* textual.

Na primeira tentativa, utilizou-se a mesma frase com a adição de pontuações enfáticas – "Estou feliz!!!!" – Na expectativa de induzir uma entonação mais animada, mesmo sem alterar os parâmetros técnicos da plataforma. No entanto, o resultado obtido[4], revelou uma limitação importante: o áudio gerado manteve uma entonação neutra e robótica, evidenciando que a modificação textual isolada não foi suficiente para acionar uma resposta emocional coerente por parte do sistema.

Para aprimorar os resultados obtidos, foram realizados ajustes nos parâmetros avançados disponibilizados pela plataforma, especialmente os controles de "estilo de exagero" e "estabilidade". A estabilidade, cujo valor padrão é 50%, foi levemente reduzida para 38%, visando aumentar a variabilidade natural da fala. Já o parâmetro de estilo de exagero, que inicialmente se encontra em 0%, foi elevado para 30%, o que contribuiu para uma entonação mais expressiva e energética. Apesar dessas modificações terem proporcionado melhorias perceptíveis[5], os áudios gerados ainda não alcançaram a fidelidade emocional presente na gravação original, que expressava claramente um estado de alegria.

Diante disso, foi conduzido um novo teste com o objetivo de intensificar a expressividade emocional da fala sintetizada. Para isso, optou-se por utilizar a palavra isolada "Alegria", escrita de forma alongada e com pontuação enfática como "AAlegria!!!", a fim de induzir uma entonação mais animada. Esse recurso, combinado com múltiplas regenerações (ou seja, diferentes tentativas automáticas de geração do mesmo trecho), resultou em versões que se aproximaram mais da expressividade esperada.

Além disso, foram ajustados o "estilo de exagero" e a "estabilidade" da plataforma. Embora a documentação da ferramenta não recomende valores elevados para o parâmetro de estilo, experimentou-se um aumento de até 50%, o que intensificou a variação de frequência e a ener-

[3] https://drive.google.com/file/d/1AD7jfwIrfg871LdfUa17FAkdTcplrk7c/view
[4] https://drive.google.com/file/d/1Gx3XBDGi1mRxJbzyJsL2rp37ULIoTlmx/view
[5] https://drive.google.com/file/d/1oACc-xhcHHQtSCUeDIgHc98WvbJTeMky/view

gia vocal na fala sintetizada. Esse ajuste contribuiu para a obtenção de resultados mais satisfatórios, com timbres que se mostraram mais coerentes com a emoção de alegria que se buscava transmitir[6].

Na tentativa de reproduzir a frase original "Estou triste"[7], observou-se um nível mais elevado de complexidade em comparação aos testes voltados à expressão de alegria. O principal desafio consistiu em obter um timbre vocal mais grave e uma entonação melancólica, características essenciais para transmitir a emoção de tristeza de forma convincente. No entanto, a plataforma utilizada não oferece controle direto sobre aspectos específicos como altura tonal ou coloração do timbre, o que limita as possibilidades de ajuste fino.

Na primeira tentativa, foi empregado o *prompt* "Ah, estou triste!"[8], com a intenção de induzir uma entonação mais emocional. Embora as respostas geradas tenham apresentado certa expressividade, a maioria soava artificial ou forçada, comprometendo a naturalidade da fala sintetizada. Diante disso, adotou-se uma nova abordagem, baseada na redução drástica do parâmetro de "estilo de exagero". Essa alteração resultou em uma fala mais contida, com menor variação prosódica, o que contribuiu para uma entrega mais sóbria e condizente com o estado emocional desejado. Ainda assim, o resultado obtido – representado pela frase "Estou triste"[9] – demonstrou que há espaço para melhorias, sobretudo no que diz respeito à naturalidade do timbre e à fidelidade da emoção transmitida.

Também foram conduzidos outros testes com a expressão original "Estou com sono"[10], com o objetivo de avaliar a capacidade da IA em representar a emoção de sonolência por meio de diferentes estratégias de *prompting*. Na primeira versão, gravada há aproximadamente oito meses, utilizou-se o mesmo *prompt* de referência para comparação, o que resultou em um áudio sem emoção. Na segunda tentativa, utilizou-se a mesma frase, mas com variações na grafia e na prosódia artificialmente induzida, como em "Estouu cum sono" – o que gerou, de forma inesperada, um áudio com entonação interrogativa.

[6] https://drive.google.com/file/d/1pRmbbBYZF7kzLlvauD0OXH1Ex_VYQ_8C/view
[7] https://drive.google.com/file/d/1FixsQ94zwUXiFX2zM71p3oE8gD9gGPjW/view
[8] https://drive.google.com/file/d/1-yHKBD7BlppJl5Qqpr182xO16hw5MEp8/view
[9] https://drive.google.com/file/d/1lnVvUHiucjCzuaPtQdMtq1pvvSjDsHX-/view
[10] https://drive.google.com/file/d/1QUOFUUhII1EyWGoG8Sychy4aMqwhnW-n/view

Em seguida, foram testadas outras variações criativas, como "aaaa estou com sonooo" e novamente "estouu cum sono", com ajustes nos parâmetros de velocidade e de similaridade vocal, com o intuito de forçar um ritmo mais arrastado e entorpecido, típico de quem está prestes a adormecer, como evidenciado na sequência de áudios[11]. No entanto, o resultado final permaneceu artificial, com um tom notavelmente robótico.

Acredita-se que essa limitação esteja relacionada à dificuldade da plataforma em reproduzir nuances fisiológicas específicas, como a vocalização anasalada e a articulação preguiçosa associadas ao bocejo – elementos cruciais para a comunicação convincente do estado de sonolência. Até o momento, os recursos disponíveis ainda não permitem simular essas características com precisão.

A partir da análise comparativa dos testes realizados com as expressões "Estou feliz", "Estou triste" e "Estou com sono", observou-se um padrão de comportamento na geração vocal da IA que evidencia uma limitação importante: a ausência de continuidade contextual entre emoções distintas. Quando uma série de tentativas é dedicada a uma emoção específica – como tristeza, por exemplo – a IA parece internalizar temporariamente certos padrões prosódicos e timbrais, o que resulta em respostas progressivamente mais coerentes com o sentimento desejado.

Isso foi percebido durante o refinamento do áudio relacionado à tristeza, em que o tom se tornou mais sóbrio à medida que ajustes como a redução do "estilo de exagero" foram aplicados. Contudo, ao iniciar um novo ciclo com uma emoção oposta, como alegria, essa memória operacional não é mantida. O sistema se comporta como se estivesse partindo do zero, ignorando aprendizados anteriores e exigindo novamente uma sequência extensa de ajustes, variações de *prompt* e múltiplas gerações.

Essa dificuldade torna-se ainda mais evidente nos testes relacionados à sonolência, em que a ausência de componentes fisiológicos – como o som anasalado característico de bocejo ou a articulação mais relaxada – compromete a naturalidade da expressão. Em contextos pedagógicos, essa limitação tem implicações diretas: a ferramenta não consegue alter-

[11] https://drive.google.com/file/d/1bGMVWH8LxpubTsbll93lQSrMmvR9YKzp/view

nar com fluidez entre diferentes estados emocionais de forma responsiva ao conteúdo ou à situação de aprendizagem. Isso impacta negativamente sua eficiência em atividades que demandam variação afetiva rápida e coerente, como narrações dramáticas, simulações de situações reais ou interações com estudantes em idade infantil, que dependem fortemente da expressividade vocal para compreender e se engajar com o conteúdo apresentado.

Ainda, no contexto do método fônico, a produção de fonemas isolados por meio de IA revelou-se um desafio considerável. A tecnologia apresentou limitações ao tentar reproduzir sons guturais, nasais e fricativos – como o fonema da letra "f". Foram realizados testes com a inserção de pontuações específicas, caracteres especiais, interjeições e letras repetidas, com o objetivo de induzir variações fonéticas – por exemplo, "f...f...fá" ou "ffffáá", na tentativa de representar o som /f/[12]. Outra estratégia adotada foi o uso de palavras naturais que contêm, em sua estrutura, os fonemas desejados.

Um exemplo analisado foi a palavra "afta", selecionada por conter o som da letra "f" de forma sutil e isolável. A proposta era que, com áudios de qualidade suficiente, fosse possível extrair trechos que reproduzem fonemas puros, viabilizando a criação de um banco de sons compatível com os princípios do método fônico. No entanto, os resultados não atenderam às expectativas: as pronúncias geradas pela IA apresentaram distorções e imprecisões que comprometeram a clareza dos fonemas, inviabilizando sua aplicação em contextos educativos.

Os testes realizados evidenciaram que a limitação das vozes sintéticas na expressão convincente de emoções continua sendo um dos principais obstáculos para sua adoção plena em contextos pedagógicos. A incapacidade de controlar de forma precisa e replicável a variação de entonação e a adaptação emocional compromete a coerência afetiva dos áudios utilizados nas interações com o robô BITI, afetando o engajamento e a eficácia comunicativa, especialmente quando se busca simular emoções humanas em atividades educativas.

É importante destacar que, embora os resultados sejam considerados adequados para o contexto de pesquisa, especialmente devido ao

[12] https://drive.google.com/file/d/1cBtdf0a-AEMSb6ySfnT8m1YoprtCX7Z3/view

baixo custo e à praticidade da IA, eles ainda estão aquém do esperado para aplicações artísticas ou profissionais. Nesses casos, a atuação de um dublador humano continua sendo superior, sobretudo na reprodução de nuances emocionais, entonações realistas e controle prosódico – aspectos que, até o momento, a voz sintética não é capaz de replicar com fidelidade comparável.

CONSIDERAÇÕES FINAIS

Este estudo representa um avanço nas investigações com a plataforma Eleven Labs ao enfrentar o desafio de aprimorar a expressividade emocional do robô BITI, contribuindo para o desenvolvimento de soluções em robótica assistiva voltadas à educação inclusiva. A pesquisa confirma achados anteriores ao demonstrar que a consistência e a previsibilidade vocal são elementos-chave para promover o engajamento de crianças com TEA em interações mediadas por robôs (Feil-seifer; Mataric, 2009; Kozima *et al.*, 2005).

Os testes com clonagem vocal por IA revelaram que, apesar dos avanços recentes, as tecnologias de síntese de fala ainda enfrentam limitações na reprodução natural de emoções humanas, sobretudo em frases que exigem maior carga afetiva. Dificuldades no controle preciso da entonação, do timbre e da prosódia emocional comprometem a fidelidade das expressões geradas, reduzindo sua aplicabilidade em contextos pedagógicos que exigem variabilidade e coerência emocional.

Conclui-se que a abordagem mais viável no cenário atual é o uso de estratégias híbridas, combinando ajustes refinados de parâmetros, variações criativas de *prompts* e curadoria humana. Para pesquisas futuras, recomenda-se o desenvolvimento de modelos específicos treinados com vozes infantis e dados emocionais, bem como a adaptação dessas tecnologias às variações regionais da língua portuguesa. Além disso, destaca-se a importância de explorar o uso funcional de diferentes vozes conforme o perfil emocional desejado, e de construir bibliotecas validadas de expressões que possam ampliar a eficácia das aplicações educacionais baseadas em síntese vocal.

REFERÊNCIAS

ABREU, M. C. F. **Desenvolvimento de conceitos científicos em crianças com deficiência mental**. 2006. 114f. Dissertação (Mestrado) – Universidade Católica de Brasília, Brasília.

ALBINO, Bruce dos Santos; SANTOS, Rafael Silva. Um levantamento bibliométrico: o uso de IA para Neurodivergentes na Educação Básica Brasileira. **Com a Palavra, o Professor**, v. 9, n. 25, p. 116-134, 2024.

BAR-SHALITA, T.; VATINE, J. J.; PARUSH, S. Sensory modulation disorder: a risk factor for participation in daily life activities. **Developmental Medicine & Child Neurology**, v. 50, n. 12, p. 932-937, 2008. Disponível em: Sci-Hub.

BRASIL. **Lei nº 12.764, de 27 de dezembro de 2012**. Institui a Política Nacional de Proteção dos Direitos da Pessoa com Transtorno do Espectro Autista; altera o § 3º do art. 98 da Lei nº 8.112, de 11 de dezembro de 1990; e cria a Semana Nacional de Conscientização sobre o Autismo.

BRASIL. **Lei nº 14.254, de 30 de novembro de 2021**. Dispõe sobre o acompanhamento integral para educandos com dislexia ou Transtorno do Déficit de Atenção com Hiperatividade (TDAH) ou outro transtorno de aprendizagem.

CARRARO, Fabrício. **Inteligência Artificial e ChatGPT:** da revolução dos modelos de IA generativa à Engenharia de Prompt. Casa do Código, 2023.

CASANOVA, Edresson *et al.* Recursos para o processamento de fala. In: **Processamento de linguagem natural:** conceitos, técnicas e aplicações em português. 2023.

CHEN, W.; JIANG, X. Voice-Cloning Artificial-Intelligence Speakers Can Also Mimic Human-Specific Vocal Expression. **Preprints**, 2023. DOI: 10.20944/preprints202312.0807.v1.

CRIVELLARO, Marcos E.; SILVA, Rogério E. da. Expressão de emoção na voz gerada por conversão texto-fala. **Anais do Computer on the Beach**, v. 3, p. 397-398, 2012.

DIAS, A. R. *et al.* Tecnologias assistivas: cartões RFID como ferramenta de auxílio na comunicação de crianças com TEA. In: ESCOLA REGIONAL DE INFORMÁTICA DE MATO GROSSO (ERI-MT), 12, 2023, Cuiabá/MT. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023.

FEIL-SEIFER, David; MATARIĆ, Maja J. **Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders**. In: EXPERIMENTAL ROBOTICS: The Eleventh International Symposium.

FRAGOSO, Analice Oliveira *et al.* Dificuldade de leitura em crianças com transtorno de déficit de atenção e hiperatividade: Relato de intervenção com método fônico. **Cadernos de Pós-Graduação em Distúrbios do Desenvolvimento**, v. 13, n. 1, 2013.

GOMES, Erissandra; PEDROSO, Fleming Salvador; WAGNER, Mário Bernardes. Hipersensibilidade auditiva no transtorno do espectro autístico. **Pró-Fono Revista de Atualização Científica**, v. 20, p. 279-284, 2008.

INTERNATIONAL PHONETIC ASSOCIATION (IPA). [S. l.], (2025). Disponível em: https://www.internationalphoneticassociation.org/. Acesso em: 7 maio 2025.

JAWORSKI, N.; HIBEAULT, M. D. Technology for teaching: Audacity. Free and open-source software. **Music Educators Journal**, v. 98, n. 2, p. 39-40. 2011. DOI:10.1177/0027432111428745. 2011.

KOZIMA, Hideki; NAKAGAWA, Cocoro; YASUDA, Yuriko. **Interactive robots for communication-care:** A case-study in autism therapy. In: IEEE INTERNATIONAL WORKSHOP ON ROBOT AND HUMAN INTERACTIVE COMMUNICATION - ROMAN 2005. IEEE, 2005. p. 341-346.

MARTTINS, L. S. **O processo de alfabetização através do método fônico**. 2022. 20 f. Trabalho de Conclusão de Curso (Especialização em Pedagogia Educação Profissional e Tecnológica) – Instituto Federal Goiano, Campus Iporá.

NEEKHARA, P. *et al.* Expressive neural voice cloning. In: ASIAN CONFERENCE ON MACHINE LEARNING, November. **Proceedings** [...]. PMLR, p. 252-267, 2021.

OLIVEIRA, J.; ALBUQUERQUE, F. E. Leitura e escrita em crianças com autismo: o trabalho psicopedagógico a partir do método fônico na Clínica Escola Mundo Autista. **Facit Business and Technology Journal**, v. 1, n. 29, p. 1-15, 2021.

PALMA, Maria Vittória S. R. *et al.* **Integração de Inteligência Artificial e Clonagem de Voz para Manter a Autenticidade e Aperfeiçoar a Interação do Robô Otto com Crianças com TEA**. In: Escola Regional de Informática de Mato Grosso (ERI-MT). SBC, 2024. p. 102-107.

PEPINO, L.; RIERA, P.; BARCHI, G.; GIACCIO, J. Sistema de conversión de texto a habla en español con control de acento, prosodia y clonación de voz. **Memorias de las JAIIO**, v. 8, n. 2, p. 110-111, 2022.

PINHEIRO, A. P. *et al.* Emotional authenticity modulates affective and social trait inferences from voices. **Philosophical Transactions of the Royal Society B**, v. 376, n. 1840, 2021. Disponível em: https://doi.org/10.1098/rstb.2020.0402. Acesso em: 7 maio 2025.

REBOUÇAS, Gabriel R. B. *et al.* O potencial da robótica no tratamento terapêutico de crianças com Transtorno do Espectro Autista. In: WORKSHOP SOBRE AS IMPLICAÇÕES DA COMPUTAÇÃO NA SOCIEDADE (WICS), 2023. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 173-183.

ROSENBERG-KIMA, Rinat *et al.* **Human-Robot-Collaboration (HRC):** social robots as teaching assistants for training activities in small groups. In: ACM/IEEE INTERNATIONAL CONFERENCE ON HUMAN-ROBOT INTERACTION, 14., 2019. IEEE, 2019. p. 522-523.

SCASSELLATI, B.; ADMONI, H.; MATARIC, M. Robots for use in autism research. **Annual review of biomedical engineering**, v. 14, p. 275-294, 2012.

SEBRA, A. G.; DIAS, N. M. Métodos de alfabetização: delimitação de procedimentos e considerações para uma prática eficaz. **Revista Psicopedagogia**, v. 28, n. 87, p. 306-320, 2011.

SHAYWITZ, Sally E. *et al.* Neural systems for compensation and persistence: young adult outcome of childhood reading disability. **Biological Psychiatry**, v. 54, n. 1, p. 25-33, 2003. Elsevier.